A HYDROPHOBICITY BASED NEURAL NETWORK METHOD FOR PREDICTING TRANSMEMBRANE SEGMENTS IN PROTEIN SEQUENCES

Zhongqiang Chen, Qi Liu, Yisheng Zhu, Yixue, Li*, Yuhong Xu
Department of Bioimedical Engineering, Shanghai Jiaotong University, Shanghai, 200030, China
Email: czqbme@yahoo.com
Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 200031, China

Abstract-**Transmembrane proteins play vital roles in living cells. The difficulties in determining the topology of transmembrane protein experimentally and the increasing amino acid sequence data from genome projects provide great demand for computational methods to predict the region of transmembrane segments in protein sequences. A hydrophobicity based supervised learning vector quantiztion neural network prediction method is presented. The prediction accuracy is above 90% and comparable to existing methods.**
Keywords- **learning vector quantization (LVQ), hydrophobicity, transmembrane(TM)**

I. INTRODUCTION

Transmembrane proteins are integral membrane proteins that span the phospholipid bilayer completely. There are two basic ways these proteins can transverse the membrane, as one or more α-helix or as a β sheet barrel. Because their functions vary from ion channel to receptor and factor, transmembrane proteins play significant and functionally distinct roles in living cells. The number of the protein sequences in the Protein Information Resource (PIR)[1] and SWISS-PROT [2] database has increased exponentially as a result of genome projects. It is estimated that 20-30% of all genes in most genomes encode membrane proteins [3]. However, due to the difficulties in purification and crystallization of these proteins and limitation of other method such as nuclear magnetic resonance (NMR), only a handful membrane proteins whose topologies have been verified by experimental methods are available [4]. Hence, there's a great demand for computational methods to predict the secondary structure of transmembrane protein.

There are several methods developed to predict transmembrane segments in membrane proteins. The earliest method is based on hydrophobicity analysis by Ktye and Doolittle [5] and later by Engelman [6]. The accuracy was improved by considering different charge distribution between the inside and outside loops [7]. As the experimental data increased, statistical approach on the amino acid distribution in various structural parts of membrane protein came into being [8]. By combining multiple sequence alignment, neural network based algorithm drawing information from aligned protein sequences reached a high accuracy above 90% [9][10]. Recently two methods using hidden markov model (HMM) also give good results [11][12]. Although hydropathy plot is still widely used, one main problem is to determine the value of cutoff on the hydropathy plot. Here we use a supervised learning vector quantization (LVQ) neural network to automatically determine the helical transmembrane region based on hydrophobicity of amino acid sequence.

II METHODOLOGY

A. Datasets

We use the latest version (released by 2001-01-29) of MPtopo database [4]. Although there are more than 1000 records of membrane protein in SWISS-PROT, most of their topology have not been verified by experimental means [13]. The reliability of transmembrane sequence assignment for those membrane proteins is insecure. MPtopo is a database of membrane proteins whose topologies have been verified experimentally by means of crystallography, gene fusion, and other methods [4]. The MPtopo database contains 547 transmembrane segments belonging to 92 proteins that were divided into 3 catalogs: 3D_helix, 1D_helix and 3D_other [4]. The length distributions of transmembrane segments in these proteins are shown in Fig.1. There are apparent differences among the distributions of three catalogs. The average length of transmembrane segments in 3D_helix is greater than that of 1D_helix and 3D_other. To avoid the bias in choosing the

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25OCT2001	N/A	-

Title and Subtitle A Hydrophobicity Based Neural Network Method for Predicting Transmembrane Segments in Protein Sequences	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Department of Bioimedical Engineering, Shanghai Jiaotong University, Shanghai, 200030, China	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on CD-ROM., The original document contains color images.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

data, we select half of the proteins in all 3 catalogs as training set for neural network, and divide the rest of the data equally into two set: validation set to evaluate the training result and test set to test the performance of the trained neural network.

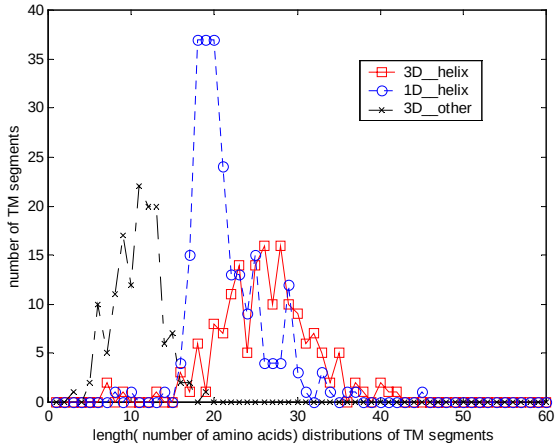

Fig.1. length distributions of transmembrane segments

B. Learning Vector Quantization

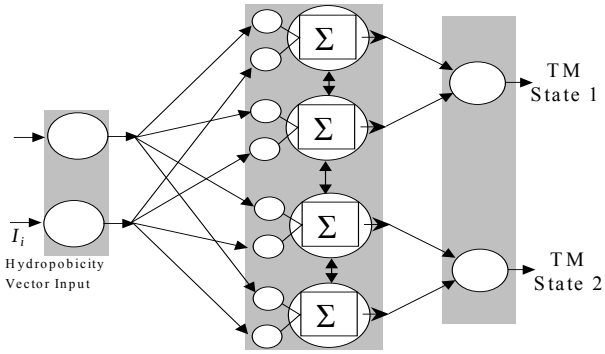

Fig.2. Learning Vector Quantization neural network architecture for hydroprobicity prediction. The hidden layer(competitive layer) use a winner-take-all mechanism. The output layer is linear layer

Learning vector quantization (LVQ) by Kohonen[14] is a supervised algorithm for training vector quantization (VQ) classifiers that in practice nearly always rapidly converge to a "good" solution. It generates a table of vector templates known as codebook vectors or reference vectors (RVs). Each codebook vector is associated with a class. During classification, the Euclidean distances between the input vector and all of the codebook vectors are computed. The input vector is assigned to the class corresponding to the nearest codebook vector so as to minimize an error function. A learning vector quantization neural network architecture with hydrophobicity input vectors is shown in Fig. 2. During network training process, the output value is one of the three states: -1(inside loop), 0(transmembrane region), +1(outside loop).

C. Training the Neural Network

We train the learning vector quantization neural network through following steps:

(1) Initialization of codebook vectors. We select equal number of codebook vectors randomly chosen from each class.

(2) Determine the nearest codebook vector. For each training vector, compute the Euclidean distance to each codebook vector,

$$d_j = \left\| w_j - I \right\| = \sum_i (w_{ji} - I_i)^2$$

Determine the codebook vector nearest the input vector,

$$c = arj \min_j (d_j)$$

(3) Learning process. In the winning processing, if an element has the same class (minimum distance) association as the training vector, modify its codebook vector close to the training vector, otherwise move it away. The other codebook vectors are left unchanged,

$$m_c(t+1) = m_c(t) + s(t)\alpha_c(t)[I(t) - m_c(t)]$$

$$\begin{cases} s(t) = +1 & \text{if } I \text{ is classified correctly} \\ s(t) = -1 & \text{if } I \text{ is classified incorrectly} \end{cases}$$

$$m_i(t+1) = m_i(t) \qquad \text{for } i \neq c.$$

where $0 < \alpha_c(t) < 1$ is a learning rate which provides fast convergence.

We use LVQ2 [15] which features on fine tuning the decision borders between the compete classes. We use the training set to train the network and stop training when the network output satisfy the validation dataset.

III RESULTS

A. Neural Network Prediction Improves Accuracy

Because the adaptiveness of the neural network, the prediction result is better than original hydrophathy plot. Using an amino acid sequence of Cytochrome Bc1 Complex from Bovine (PIR identity: CBBO) as input, both of the results of hydrophobicity plot and the neural network prediction output are shown in Fig.3.

From the figure we can see that it is difficult to correctly determine the transmembrane segments from the hydropathy plot. Especially the 2nd and 7th peaks in the upper part of the figure are much lower than other hydropathy peaks. Usually people need to specify the low cutoff and high cutoff on the hydropathy plot. The output of hydrophobicity based LVQ neural network need not any preseting about the cut-off and the result is fairly good according the true transmembrane segments. Because the 2nd and 3rd segments are too close, the neural network recognize them as one segment (Fig.2. lower part)

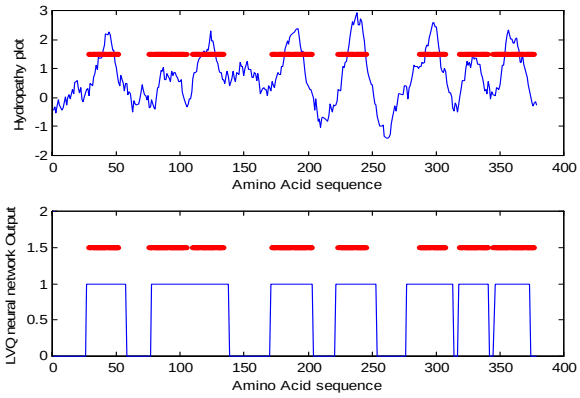

Fig. 3. Comparison between hydropathy plot (the upper part) and hydrophobicity based LVQ neural network prediction output(the lower part).

B. Test Data Set Evaluation

We use the test set that is totally independent to the training set and validation set. The result is show in Table 1.

The accuracy is high and comparable to that of existed method. This is partly due to the small number of records in test set. There are 6 segments have been predicted wrongly. Most of the segments have been successfully predicted. The single transmembrane helix sensitivity ($M=N_{correct}/N_{known}$) reaches 94% and the single transmembrane helix specificity($C=N_{correct}/N_{predicted}$) reaches 96%. The value M is unexpectedly low, we will discuss it in next section. Since we haven't got larger test set yet, we are unable to compare the same large data set to the other methods.

TABLE 1
Prediction accuracy of different algorithm using test dataset

Algorithm	Number of transmembrane helices[a]			Q(%)[b]
	N_{known}	$N_{correct}$	$N_{predicted}$	
PHDhtm[c]	145	142	144	96
TMHMM[d]	145	140	143	95
LVQ[e]	145	137	143	91

[a] N_{known}, $N_{predicted}$, $N_{correct}$ are, respectively, number of experimentally known helices, total number of predicted, and number predicted correctly. $N_{correct}$ is defined as predicted helices that exhibited at least a 50% overlap with known transmembrane helices.

[b] Prediction accuracy Q was determined as described in Tusnády and Simon (1998).

$$Q = 100\sqrt{\frac{N_{correct}}{N_{known}}\frac{N_{correct}}{N_{predicted}}}.$$

[c] From the PredictProtein automatic prediction server [9] [10] using the default settings.

[d] Hidden Markov Model [12](TMHMM) used with single sequence information from MPtopo.

[e] Hydrophobicity based learning vector quantization using Kyte and Doolittle's hydrophobicity scale, window size=19.

IV. DISCUSSION

Different hydrophobicity scale may lead to different result. We use the most widely used Kyte and Doolittle's hydrophobicity scale in this paper. We have tried Engelman's hydrophobicity scale, but the result is not so good as Kyte and Doolittle's. Although we haven't tried more scales, other scale may be a good choice.

Compared to the existing neural network methods, which use multiple sequence alignment result and require both large memory and computational time, our LVQ neural network require small memory and can be trained within minutes. For overall topology prediction, the prediction results of our method is however not as good as PHDhtm.

The neural network architecture itself doesn't provide further biological meaning. In this sense, hidden markov model method is better than our neural network, since it

architecture follows the biological model to some extend.

In many cases, two adjacent transmembrane segments are so close that the neural network can't separate one from another. Consequently, there are merged transmembrane segments in the prediction result. These lead to decrease in the number of helices predicted correctly and drops in the overall accuracy(Q), especially in the single transmembrane helix sensitivity ($M=N_{correct}/N_{known}$). Hence, the parameters of this neural network model presented in this paper need further tuning.

V. CONCLUSION

In this paper we present a hydrophobicity based neural network prediction method, which successfully predict the regions of transmembrane segments in proteins. The overall accuracy is high and comparable to other methods. LVQ neural network shows its power to automatically determine the threshold of hydrophobicity value of transmembrane protein. This indicates that there is room for other automatic methods in determination of transmembrane segments in protein.

ACKNOWLEDGEMENTS

We thank Prof. Stephen White for the MPtopo database and his helpfull discussion.

REFERENCE

[1] Baker, W.C., Garavelli, J.S., Huang, H.Z., McGravey, P.B., Orcutt, B.C., Srinivasarao, G.Y., Xiao, C.L., Yeh, L.-S.L., Ledley, R.S., Janda, J.F., et al, "The Protein Information Resource(PIR)," *Nucleic Acids Res.*, vol.28, pp.41-44, 2000.

[2] Bairoch, A. and Boeckmann, B., "The SWISS-PROT protein sequence data bank," *Nucleic Acid Res.*, vol. 19, pp.2247-2248, 1991.

[3] Krogh A., Larsson B., von Heijne G., Sonnhammer E.L., "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *J. Mol. Biol.*, Vol.305(3), pp.567-580, 2001.

[4]Jayasinghe, S., Hristova, K., White, S.H., "Mptopo: A database of membrane protein topology," *Protein Sci.*, vol.10, pp.455-458, 2000.

[5] Kyte, J., Doolittle, R.F., "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, Vol.157(1), pp.105-132, 1982.

[6] Engleman, D.M. Steitz, T.A. and Goldman, A. "Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins," *Annu. Rev. Biophys. Chem.*, vol.15, pp.321-353, 1986.

[7] von Heijine, G., "The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology," *EMBO J.*, vol.5(11), pp.3021-3027, 1986.

[8] Jones, D.T., Taylor, W.R. and Thorton, J.M.,"A model recognition approach to the prediction of all helical membrane protein structure and topology," *Biochem.*, vol.33, pp.3038-3049, 1994.

[9] Rost, B., Casadio, R., Fariselli, P. and Sander, C., "Transmembrane helices predicted at 95% accuracy," *Protein Sci.*, vol.4(3), pp.521-533, 1995.

[10] Rost, B., Fariselli, P. and Casadio, R., "Topology prediction for helical transmembrane proteins at 86% accuracy," *Protein Sci*., vol.5(8), pp.1704-1718, 1996.

[11] Tusnady, G.E., Simon I., "Principles governing amino acid composition of integral membrane proteins: application to topology prediction," *J. Mol. Biol.*, vol.283(2), pp.489-506, 1998.

[12]Sonnhammer, E.L., von Heijne, G., Krogh, A.A. "Hidden Markov model for predicting transmembrane helices in protein sequences," *Proc. 6th Int. Conf. Intell. Syst. Mol. Biol.*, pp.175-182, 1998.

[13]Senes, A., Gerstein, M., and Engelman, D.M., "Statistical analysis of amino acid patterns in transmembrane helices: The GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions," *J. Mol. Biol.,* vol. **296,** pp.921–936, 2000.

[14]Kohonen, T., "The self-organizing map", *Proceedings of the IEEE*, vol.78(9), pp.1464-1480, 1990.

[15] Kohonen, T., Hynninen, J., Kangas, J., Laakosonen, J. , and Torkkola, K., "LVQ_PAK: The learning vector quantization program package" Helsinki University of Technology, Laboratory of Computer and Information Science Technical Report A30, 1996.